Unmasking An Unseen Influence:
Trading by Uninformed Retail Investors and Its Capital Market Effects

By

Stephen P. Ferris
Indiana University of Pennsylvania
Email: ferris@iup.edu

Jan Hanousek, Jr.
University of Memphis
Email: jan.hanousek@memphis.edu

Jan Hanousek
Mendel University in Brno, and CEPR London
Email: jan.hanousek@mendelu.cz

Jolana Stejskalová
Mendel University in Brno
Email: jolana.stejskalova@mendelu.cz

January 2024

We are grateful for the thoughtful comments of Konstantin Sokolov on earlier versions of this manuscript.

Unmasking An Unseen Influence:
Trading by Uninformed Retail Investors and Its Capital Market Effects

Abstract

Using a natural experiment based on technical improvements to Google Trends data, we can more clearly separate less informed from more informed retail investors. We find that uninformed trading has a significant negative effect on liquidity, although the effect is most pronounced for smaller firms. This adverse effect of uninformed trading also increases the cost of capital for smaller firms. Our results help to explain the mixed evidence in the literature regarding the effect of uninformed trading on market liquidity.

JEL Codes: G12; G14

Keywords: uninformed, retail investors, liquidity, trading, cost of capital

Unmasking An Unseen Influence:
Trading by Uninformed Retail Investors and Its Capital Market Effects

1. Introduction

What effect do retail investors have on the equity capital market? Researchers such as Frazzini and Lamont (2008), Hvidkjaer (2008), and Barber et al., (2009) view them as uninformed. Shiller (1984) argues they are simply noise traders, providing no useful news or information. But others such as Kumar and Lee (2006), Kaniel et al. (2012), and Barrot et al. (2016) contend that their participation has an overall positive effect since they provide liquidity to the market. This is the liquidity provision hypothesis of retail trading.

However, the claim that retail investors provide market liquidity is inconsistent with more recent findings. Bradley et al. (2023) suggest that retail investors have coordinated their trading, Goldstein et al. (2013) theorize that they propagate trading frenzies, while Chapkovski et al (2023) contend that they are motivated by emotions and gamified platforms. Further, the trading volume accounted for by retail investors has increased, now accounting for 20% of the market's volume (McCabe, 2021). Given these more recent findings regarding the effect of retail investors, the volume of their trading, and the variation in informativeness among them (Farrell et al., 2022), we believe that their effect on the equity market extends beyond providing liquidity.

To undertake our empirical analysis, we must distinguish between the attention of less-informed and more-informed retail investors. By using technological improvements in the construction of the Google Search Volume Intensity as a natural experiment, we separately measure the attention of less informed and more informed investors. This approach allows us to test whether less informed retail investors are really liquidity providers and whether their participation affects the cost of capital for publicly traded firms.

Research by scholars such as Hirshleifer and Teoh (2003), Hou et al. (2009), and Hirshleifer et al. (2011) contend that investor attention is necessary for a stock price to fully reflect public information since investors need to be aware of the information before they can react to it. Ding and Hou (2015) show that Google search data captures the information-seeking activity of investors. As a result, Google search data allows us to study the impact of investor attention on liquidity.

Google provides aggregated data for the intensity of firm-level searches using its platform, Google Trends. Ding and Hou (2015) show that this data captures investor attention in a specific stock. This data is, however, continually updated by Google, even retrospectively, to make the data more precise.[1] Google announced three significant technical improvements in the data collection and classification for the Google Trends data. [2] We use the latest improvement in January of 2022 as a natural experiment for distinguishing between less and more informed retail investors. We download the Search Volume Intensity (SVI) of the tickers for all publicly listed firms from 2004 to 2018, both before and after the improvement. By subtracting the SVI of the newer dataset from the SVI of the older dataset values, we can observe searches that Google categorizes as unrelated. These types of searches can proxy for the searches of the less informed traders since their searchers are more likely to be identified as irrelevant/unrelated by Google.

We find that the attention of investors in the aggregate increases liquidity, but that the attention of less informed investors has the opposite effect. As noted previously, these investors

[1] For example, Google has made over 5,000 improvements to their searches in 2021 alone (https://blog.google/products/search/danny-25-years-of-search/).

[2] The improvements occurred in January 2011, January 2016, and January 2021. Each one of these improvements introduced new features to Google Trends as well as improving the accuracy of the data.

are more likely to fall victim to herding and trading frenzies. Thus, they are less likely to make contrarian trades and unable to provide liquidity (Kaniel et al., 2008; Kelley and Tetlock, 2013). This result, however, is primarily observed for smaller firms, where individual investors constitute a larger market share. We further verify that retail investors drive this result using Barber et al.'s (2023) improvement on Boehmer et al.'s (2021) algorithm.

Overall, this adverse effect on the liquidity of smaller firm stocks by uninformed traders results in reduced performance for these same firms. The poorer performance of these firms implies a higher cost of capital since it is more difficult for such firms to attract investors. Our results are consistent with the liquidity provision hypothesis for the stocks of smaller firms.

This study offers a variety of contributions to trading, market efficiency, and microstructure literature. We introduce a new measure for identifying uninformed trading through a natural experiment made possible by Google's technical improvement of its search volume intensity factor. This identification strategy for uninformed trading can be used for a number of future studies examining issues as varied as stakeholder/shareholder conflicts, the desirability of controversial corporate policies, corporate restructuring initiatives, and the quality and nature of corporate governance. We also examine more comprehensively the effect that retail trading has on the equity market beyond liquidity to include the firm's cost of capital.

We organize this study into six sections. In the following section, we review the possible effects that equity trading by retail investors might exert in the capital market on the firm's own financial circumstances. Section three explains our data and describes how we construct our novel measure of uninformed trading by use of Google Trends data. We present our empirical findings regarding liquidity and cost of capital in section four. Section five contains our robustness tests. We summarize our findings and discuss their importance in section six.

2. The Consequences of Equity Trading by Retail Investors

2.1 Liquidity Effects

Previous literature views retail investors as noise traders (Black, 1986; Shleifer and Summers, 1990). That is, they are traders who drive prices away from value fundamentals and destabilize markets. More recent literature, however, suggests that trading by retail investors provides liquidity to other market participants (Kaniel et al., 2012; Kaniel, et al., 2008; Kelley and Tetlock, 2013). Barrot et al. (2016) show that retail investors provide liquidity, especially when conventional liquidity providers are constrained. However, recent market activity, such as GameStop's short squeeze, demonstrates the effect that retail investors can have on the market and have been accused of coordination (Bradley et al., 2023). Retail investors can also cause trading frenzies (Goldstein, et al., 2013; Bradley et al., 2023), which can affect the market's liquidity.

A potential explanation of this conflict on the effect on the liquidity of retail investor participation resides in the different levels of investor informativeness and sophistication. Retail investors' current level of participation is unprecedented, with retail volume now accounting for 20% of stock market activity (McCabe, 2021). Explanations for this rise in participation include the COVID-19 pandemic (Ozik, Sadka, and Shen, 2021), the resulting changes in work patterns, and the increase in the gamification of trading platforms. Chapkovski et al. (2023) find that gamified platforms have an impact on investor behavior and significantly increase trading volume. They also suggest that gamified platforms attract specific types of users, noting that individuals with lower financial literacy scores tend to prefer gamified platforms[3]. This suggests that the rise in retail participation might be driven by a higher participation rate among the less sophisticated or uninformed investors. Chapkovski et al (2023) further find that investors preferring gamified

[3] Platforms employing more hedonic gamification, such as confetti and achievement badges.

platforms exhibit greater behavioral biases and introduce more noise in their trading. Conversely, sophisticated investors tend to be contrarian, implying that they are more likely to be liquidity providers (Kaniel et al., 2008; Kelley and Tetlock, 2013). Thus, we hypothesize that unsophisticated or uninformed investors will demand liquidity during times when conventional liquidity providers are constrained and, therefore, negatively affect the market.[4]

It is doubtful, however, that retail investors can impact the entire market. Kumar and Lee (2006) note that even in the presence of systematic noise trading, which pushes prices away from value fundamentals, the activities of rational arbitrageurs can offset this behavior (e.g., Shiller, 1984; Shleifer, 2000; Lee, 2001). However, not every stock will have a sufficient number of institutional investors who can serve as arbitrageurs to offset the value departures due to noise trading. Indeed, small firms will have lower analyst coverage and much fewer institutional investors (O'Brien and Bhushan, 1990), making arbitrage corrections less likely. Consequently, the liquidity of small firms is likely to be negatively affected by the trading of unsophisticated and uninformed investors while that of large firms is likely to be unaffected.

2.2 Cost of Capital Effects

Liquidity can affect the financial operations of the firm, namely its cost of capital. Diamond and Verrecchia (1991) show that increased liquidity attracts large investors, which can reduce the firm's cost of capital. If unsophisticated or uninformed investors decrease the liquidity of a firm's equity, then that firm becomes less attractive to large investors. This, in turn, is likely to increase the firm's cost of capital. Consequently, we hypothesize that the adverse effect of retail investor

[4] Potential example of a situation where conventional liquidity providers are constrained might be a trading frenzy. As Goldstein et al. (2013) show, when speculators place a large weight on a common noise in information caused by a rumor, for example, it can lead to a trading frenzy. During the frenzy, all speculators wish to trade like others, which leads to significant pressure on prices. This might cause market makers to perceive themselves as uninformed and decrease liquidity in the market (Green and Smart, 1999).

trading on the firm's cost of capital will be localized for the smaller firms with lower institutional investor participation.

3. Data and the Measurement of Investor Informativeness

3.1 Google's Search Value Intensity

In this study, we use Google Trends data for the ticker searches of all publicly listed firms from 2004 to 2018. The SVI reported by Google ranges from 0 to 100, where 100 corresponds to the period of the highest search intensity for the given term. However, Google data is not constant. Eichenauer et al. (2021) provide a detailed study into the consistency of Google search data. They observe that higher frequency searches and regional search data for the same period can vary significantly across collection dates. There are two reasons for this variance. Firstly, Google does not report the total number of searches; instead, it provides an index created from a random sampling of their data. This random sampling can cause significant deviations for small population samples, such as regional search data or higher frequency data (e.g., daily). This should not cause significant deviations for large population samples.

The second reason is technological improvement in the quality of the data. Per Google, there have been numerous[5] improvements over the years, focusing on improving the data and filtering out unrelated searches. Google continually improves its Search Volume Intensity (SVI) measure (Eichenauer et al., 2021) and applies these improvements backwards to existing data. Thus, the time-series values are retroactively changed based on these technical enhancements applied by Google.

[5] For example, Google has made over 5,000 improvements to their searches in 2021 alone (https://blog.google/products/search/danny-25-years-of-search/).

But it is exactly because Google retroactively re-estimates the time series of the SVI values after each technical improvement that we have a natural experiment. In our study, we use the SVI values for all publicly listed firms from 2004 to 2018 collected at two different points in time. The first sample, referred to as the “original”, was collected in December of 2019. These values are estimated prior to the last technical improvement[6] in 2021. Our second sample, referred to as “improved,” was collected in December of 2022.[7] Thus, we have SVI data for a sample firm at the same point in time but calculated using two different methods. Again, we want to emphasize that these changes are applied retroactively, meaning that any data collected since 2021, regardless of how far back in the past the sample is constructed, the data will consist of these improved SVI measures. We plot the differences in SVI between the original and improved values in Figure 1.

------------------------------ Insert Figure 1 around here ------------------------------

Figure 1 presents the SVI for the ticker of Google, i.e., “GOOG”[8], for both the original and improved sample. From our analysis, we observe that the original sample SVI tends to be higher[9] than the improved SVI, supporting Google’s claim that it improves its measurement of SVI by filtering out the searches they consider unrelated. We argue that the searches classified by Google as unrelated or irrelevant are those initiated by unsophisticated or uninformed investors.

[6] This improvement in 2021 consisted of introducing a new feature in Trends called “Spikes,” highlighting sudden increases in search volume for a particular query. However, Google also notes that it also changed its algorithm to filter out spam and irrelevant results more effectively.

[7] We have collected the search data for the same period several times to test the robustness of our results. As noted by Eichenauer et al. (2021), we observe slight differences in the datasets, however, the differences are not significant and close to zero for datasets downloaded close to each other. Our results remain statistically identical when using data collected at different times for the improved sample, supporting the conclusion that technological improvements drive the differences in original and improved datasets. We offer a more detailed explanation in Section 5.

[8] The search is not case-sensitive. Moreover, Google Trends offers search data on both “GOOG” as a term and as a topic, with the latter also including searches such as “GOOG price.” However, Google does not share all searches they include for any given topic, meaning that using “GOOG” as a topic might include searches that investors would not search. As a result, we focus on searches of tickers as a term consistent with past literature (e.g., Da et al., 2011).

[9] It is possible for the improved SVI to be larger than the original SVI, as is visible in the picture. This may be caused by a large spike in the searches caused by unexpected major news concerning the company that was misclassified by Google algorithm as unrelated previously. However, we observe that majority of the time (for roughly 70% of the sample) SVI of the original is larger than or equal equal to SVI of the improved dataset.

3.2 Measuring Uninformed Trading

The $SVI_{i,t}$ of stock i in month t is composed of searches by both sophisticated ($S_{i,t}$) and unsophisticated or uninformed investors ($U_{i,t}$), allowing SVI to be decomposed into its components as below:[10]

$$SVI_{i,t} = U_{i,t} + S_{i,t} \tag{1}$$

We now assume that due to technological improvements, the improved SVI* should be lower or equal to SVI. This occurs because of the removal of unrelated searches originating from unsophisticated investors. That is $U^{*}_{i,t} \leq U_{i,t}$ where $U^{*}_{i,t}$ represents the searches initated by uninformed investors that remain in the SVI measure after the 2021 technical improvement implemented by Google. Consequently, $SVI^{*}_{i,t}$, the SVI calculated after the 2021 improvements can be expressed as:

$$SVI^{*}_{i,t} = U^{*}_{i,t} + S_{i,t}, \text{ where } U^{*}_{i,t} \leq U_{i,t} \tag{2}$$

Thus, the difference between SVI and SVI* is:

$$SVI_{i,t} - SVI^{*}_{i,t} = U_{i,t} - U^{*}_{i,t} + S_{i,t} - S_{i,t} \tag{3}$$

$$SVI_{i,t} - SVI^{*}_{i,t} = U_{i,t} - U^{*}_{i,t} \tag{4}$$

$$SVI_{i,t} - SVI^{*}_{i,t} = U^{\dagger}_{i,t} \tag{5}$$

We represent the difference between SVI and SVI* as $U^{\dagger}_{i,t}$, which proxies for the attention of unsophisticated or uninformed investors.[11] Creation of this proxy allows us to more carefully

[10] Ding and Hou (2015) shows that investors are more likely to use tickers to find information about the company, whereas consumers generally use the company's name. As a result, we assume that the majority of the searches of tickers will originate from investors.

[11] It is important to note that it is not the attention of all unsophisticated investors, only those identified by Google during the latest major technological improvement. However, $U^{\dagger}_{i,t}$ should have very low type one error, meaning that

examine how retail investors, especially the uninformed, affects capital market liquidity and the firm's cost of capital.[12]

Boehmer et al. (2021) test the liquidity provision hypothesis, and their results conflicted with the results of Kaniel et al. (2008). Kaniel et al. (2008) find that investors exhibit contrarian behavior, thus providing liquidity for the market, whereas Boehmer et al. (2021) find the opposite relation. A potential explanation of this difference, as offered by Boehmer et al. (2021), is the difference in the retail order imbalance variable. Kaniel et al. (2008) use marketable and nonmarketable orders in their calculations, where marketable orders are more likely to be aggressive. Kelley and Tetlock (2013) show that aggressive market orders are more likely to be informed and predict future news. Comparatively, passive orders provide liquidity rather than being informative. Boehmer et al. (2021) find only mixed support for the liquidity provision hypothesis.

Our study can complement this stream of literature by offering another explanation of the differences, namely investor sophistication, and informativeness. By being able to observe the attention of unsophisticated investors ($U_{i,t}^{\dagger}$), we can provide further tests of the liquidity provision hypothesis. The difference in the level of sophistication has been used in the model of Grossman and Stiglitz (1980), who include both informed and uninformed investors. However, all individuals in their model are ex-ante identical. The only difference between informed and uninformed is whether they have obtained information. However, this assumption comes at odds with the study by Chapkovski et al. (2023), who observe fundamental differences between investors who prefer

attention captured should stem from unrelated searches, as any differences stemming from sampling should be minimal.

[12] It is important to note that since SVI is normalized by the maximum value, which can change during the improvement, the economic interpretation can be difficult. We further conduct further tests focusing on abnormal attention, rather than the level, in Section 5. Our results remain robust.

gamified platforms and exhibit more behavioral biases and those who do not. The recent events on the stock market, including GameStop short-squeeze and so-called “meme stock mania”, can thus be explained by a higher rate of participation by unsophisticated or uninformed investors. This further lends credence to the idea that the liquidity provision hypothesis depends on retail investors’ overall level of sophistication.

Our starting sample consists of 4518 unique firms from 2010[13] to 2018. Following[14] Da et al. (2011), we collect the monthly SVI of the firm’s tickers instead of its name. Similarly, we remove firms whose tickers are single or double alphabets (e.g., “C” for Citi group) as well as firms whose tickers have generic meanings (e.g., “DO” for Diamond Offshore Drilling). Overall, our finished sample consists of 288,793 firm-month observations. Summary statistics of the sample are available in Appendix B.

4. Empirical Findings

4.1 Search Volume Intensity and Retail Volume

In this section, we present the results of our empirical analysis. First, it is essential to show that the Search Volume Intensity (SVI) is related to retail volume and thus can be used to examine the effect of retail trading on liquidity and the cost of capital. To measure retail volume, we use the Barber et al. (2023) improvement on the Boehmer et al. (2021) algorithm, which uses the Lee

[13] Google Trends data started in 2004, and we have data for searches from 2004 to 2018. However, since we analyze the relationship between SVI and retail volume, we restrict the sample to start in 2010 to improve the accuracy of the algorithm of Boehmer et al. (2021). The rest of our results are robust to the sample reduction, and results using a total sample are available upon request.

[14] Da et al. (2011) further show that while the level of SVI can be used, they prefer the change in levels of SVI, using the past six months of data. However, base SVI is preferred in our study since our primary variable is the difference between SVI and SVI*. Using change instead of base SVI could eliminate the technological improvement in SVI* and thus reduce the power and effectiveness of our tests.

and Ready (1991) quote midpoint signing method.[15] Using this algorithm, we can approximate, with a low type one error, the monthly retail volume for each stock.

From the above, we can construct two variables of interest. The first variable we denote as *Retail Volume scaled* is defined as the monthly retail volume scaled by total share volume. The second variable we denote as *Abnormal Retail volume* is the percentage change in the ratio of retail volume to the average retail volume for the stock over the past three months. For control variables, we use *size*, defined as the market value of equity; *book-to-market ratio; illiquidity*, defined by Amihud (2002); *past returns* described by Brennan et al. (2012) as well as firm and year fixed effects. Detailed definitions of these control variables are provided in Appendix A.

------------------------------ Insert Table 1 around here ------------------------------

The results of this analysis are presented in Table 1. Columns 1-3 use *Retail Volume scaled* as a dependent variable, while columns 4-6 use *Abnormal Retail volume*. We see that SVI, both from the original and improved sample, is a significant predictor of retail volume. This confirms the conclusions of Ding and Hou (2015). Moreover, our SVI difference is also a significant predictor of retail volume. These results show that SVI is significantly and positively related to retail volume, thus supporting our methodological approach for the study of uninformed investors.

4.1 Market Liquidity Effects

The primary focus of our empirical analysis is the effect of uninformed trading on a firm's equity liquidity and, subsequently, its cost of capital. To test the liquidity provision hypothesis, we use effective spread as a measure of liquidity (Chordia, Roll, and Subrahmanyam, 2001). Similar to Fang et al. (2009), we calculate the monthly effective spreadby taking the average daily effective

[15] Barber et al. (2023) suggest that this improvement yields high and homogenous accuracy rates across all stocks. See Barber et al. (2023) for a more detailed explanation of the algorithm.

spread for the given month.[16] Since the effective spread does not follow a normal distribution, we use its natural logarithm transformation for use in our regression analysis. We find that this measure of effective spread is negatively related to market liquidity, with larger positive values indicating worsening liquidity.

------------------------------ Insert Table 2 around here ------------------------------

We report the results of our analysis in Table 2. We use the same set of control variables in Table 2 as those included in Table 1. We observe that while both the original and improved SVI lead to increased liquidity, the SVI difference significantly reduces liquidity. These findings suggest that unsophisticated retail investors do not improve liquidity. This result helps to explain the mixed evidence that exists for the liquidity provision hypothesis. While retail participation positively impacts liquidity, unsophisticated or uninformed investors are likely to reinforce any liquidity shortages rather than correct them (e.g., Goldstein et al., 2013).

However, Kumar and Lee (2006) observe that institutional investors should act as rational arbitrageurs and offset any adverse impact on market liquidity from uninformed trading. Therefore, we contend that the effectiveness of rational arbitrageurs in the market will depend on the participation level of institutional investors. For large firms with significantly more institutional investors, the retail volume will be dwarfed by that of institutions. We should, therefore, observe that unsophisticated retail trading will fail to decrease liquidity for the largest stocks.

To test this hypothesis, we interact the SVI difference, our proxy for the attention of unsophisticated or uninformed investors, with size quartiles. The results of this analysis are

[16] We use dollar-weighted effective spread scaled by midquote. Results are identical when using share-weighted effective spread instead.

presented in column 4. The smallest firms constitute the base category. We can see that the smallest firms are adversely affected by unsophisticated retail investors and have significantly worse liquidity (Shleifer and Vishny, 1997). Larger firms, however, are either unaffected or have increased liquidity, consistent with Shiller (1984).

4.2 Cost of Capital

Diamond and Verrecchia (1991) show that liquidity can attract larger investors, making a firm's equity more attractive and thus lowering a firm's cost of capital. If unsophisticated or uninformed investors negatively affect the equity liquidity of smaller firms, those firms are less able to attract institutional investors and thus struggle to raise funds. Consequently, they will suffer from a higher cost of capital.

To test the impact of unsophisticated investor attention on the cost of capital, we use an indirect approach. We first relate investor attention to firm performance. This connection is critical since firms that suffer from poor performance are less able to attract buyers of their securities. This results in a decline in equity values and raises the firm's cost of capital. Thus, there is an inverse relation between the firm's performance and its cost of capital.

To undertake our performance analysis, we use the approach of Armstrong et al. (2010). Specifically, we construct twenty-five (5x5) equal-weighted portfolios for each month based on two-dimensional dependent sorts described below. We then compute one month ahead of buy-and-hold returns for each portfolio. To construct our portfolios, firms are first ranked into quintiles based on unsophisticated retail investor attention, which is the difference between the original and improved SVI values. Then, each of these five unsophisticated retail investors' attention portfolios is sorted into five size-based portfolios, resulting in twenty-five different portfolios. We then use the five-factor Fama and French (2015) model to evaluate the performance of each portfolio.

We use the excess return as our measure of performance. As noted above, our use of performance serves as an implied albeit inverse measure for the firm's cost of capital. Similar to Armstrong et al. (2010), the focus of this study is smaller firms, which typically have low coverage by analysts. Consequently, alternative specifications for an implied cost of capital, which use analyst expectations, cannot be calculated for these firms.[17]

------------------------------ Insert Table 3 Panel A around here ------------------------------

We report the results of our portfolio analysis in Table 3. In Panel A, we show the results for the smallest pentile of firm size. We only provide the results for the lowest and highest pentile of unsophisticated investor attention for brevity and ease of interpretation. While both pentiles have significant and negative alphas, the pentile with the highest attention performs significantly worse. We also present the results from the arbitrage portfolio, where we purchase the pentile with the lowest level of attention and short the pentile with the highest level of attention. This arbitrage portfolio shows that unsophisticated retail investor attention leads to significantly worse performance. This is consistent with an increased cost of capital for these firms.

------------------------------ Insert Table 3 Panel B around here ------------------------------

We show the results for the pentile of the largest firms in Panel B. With this analysis, we obtain the opposite effect. That is, the alpha is significantly positive, implying a lower cost of capital for firms with larger attention. Our results for the arbitrage portfolio are similarly consistent.

Overall, these results show that rational arbitrageurs will offset the noise trading of unsophisticated or uninformed investors, as suggested by Kumar and Lee (2006). Nevertheless, for small firms, which are less traded by institutional investors, unsophisticated investors erode

[17] Armstrong et al. (2010) offer detailed discussion and tests of this approach.

liquidity and drive equity prices from their value fundamentals (e.g., Shleifer and Vishny, 1997). This has the effect of increasing the cost of capital to these firms.

5. Robustness tests

The previous section documents that the SVI difference, our proxy for unsophisticated or uninformed investors, leads to higher illiquidity and consequently to a higher cost of capital for smaller firms. In this section, we report the findings from a set of robustness tests to verify the causality of our findings as well as show that changes in SVI are driven by technological improvements in its measurement rather than other factors such as sample bias.

5.1 Causality interpretation

To analyze the causality of our findings, we use the the potential outcome framework of the Rubin Causal Model (Holland, 1986). This model is based on two outcomes, one with and one without treatment:

$$y_{0i} = \mu_0 + \varepsilon_{0i} \text{ and } y_{1i} = \mu_1 + \varepsilon_{1i} \tag{6}$$

Formally, the model can be written as $y_{Ti} = \mu_T + \varepsilon_{Ti}$, where subscript T=1 denotes the treatment, and T=0 represents the control group. We observe only one outcome for each i, either y_{0i} or y_{1i} and the counterfactual outcomes must be estimated. We use established Randomized Control Trial (RCT) techniques to estimate the Average Treatment Effect on the Treated (ATET), where y_{0i} is estimated using the nearest-neighbor approach with an extensive set of controls.[18]

In our analysis, the outcome variable is the measure of illiquidity, i.e., effective spread. The terms μ_1 and μ_2 represent the indicators of whether there is a high level of attention from unsophisticated or uninformed investors while controlling for various characteristics. We perform

[18] The treatment effect $E[y_{1i} - y_{0i}]$ is under random assignment equal to $\mu_1 - \mu_0$, motivating our choice of a randomized control trial (RCT).

exact matching on year, month, and the Fama-French 48 industry classification, while the approximate coordinate for matching is firm size. We follow established procedures for constructing the control group (e.g., Rosenbaum and Rubin, 1985; Rubin, 2008) and evaluating the average treatment effect on treated (ATET). To define the treatment and control group, we use the distribution of the SVI difference. Specifically, we assign monthly values above the 70th percentile[19] as the treatment group (i.e., attention driven by the unsophisticated investors) while values below the 30th percentile serve a control group. We report the results of our analysis in Panel A of Table 4.

------------------------------ Insert Table 4 Panel A around here ------------------------------

Column (1) uses the dollar-weighted effective spread scaled by midquote, while column (2) uses the share-weighted effective spread. We see in both columns that the treatment effect is highly significant, indicating an approximate 8% increase in the effective spread following high attention from unsophisticated or uninformed investors. As noted in the previous section, the economic interpretation of our results is difficult since our methodology does not allow us to identify the attention of all unsophisticated investors. The ATET approach, however, allows us to test the effect of unsophisticated investor attention more directly, by comparing matched samples. The result is highly significant and larger than the result estimated in Table 2. This suggests that the overall impact of unsophisticated or uninformed investors is more economically significant than suggested by the findings in Table 2. Overall, Panel A confirms our causal interpretation of the results in Section 4.1.

[19] We omit the middle 40% to better isolate the effects of attention of unsophisticated or uninformed investors. This split was also chosen to also offer a balancing of covariates used for matching. Results using other splits of the sample are available upon request.

We offer the mean differences and variance ratio for matched samples in Panel A. We show the balance plot for firm size, our approximate coordinate for matching, in Panel B. Overall, these results show that our sample is well balanced and supports our causal interepretation.

5.2 Placebo tests

In Section 3, we describe the data as well as the differences in the SVI provided by Google. We argue that the changes in SVI are due to technological improvements, consistent with the observation that Google improves its algorithms and applies any changes retroactively. Notably, Eichenauer et al. (2021) test the consistency of Google Trends data and find that small population data and higher frequency data may suffer from significant sample bias. While this bias should not be significant for large populations and the monthly data that we use in this study, there are still concerns that our results may be driven by sample bias rather than these technological improvements.

If our results are driven by sample bias, it would mean that the difference in SVI is purely random, lacking any systematic component. We, therefore, use a placebo test to randomly assign some firms with uninformed investor attention. If this placebo test does not lead to a significant effect on the dependent variable, it will reject the hypothesis that the difference in SVI is random[20]. Thus, the differences in SVI would be systematic, and the explanation would be the technological improvement by Google in the measurement of their data. Such results would thus verify our methodology for the identification of unsophisticated or uninformed retail investors.

[20] An Alternative test of whether the SVI difference is caused by sample bias could by done by using SVI data downloaded at different times both prior and post the technological improvements. While we repeated our tests with SVI downloaded at later times, we only have one SVI data collected prior to the last technological advancements. Due to the nature of a placebo test and its robustness, that approach should be considered superior. Nevertheless, we replicate the analysis using SVI downloaded at different times post the 2021 improvement. The results are very similar and lead to same conclusion. For reasons of reporting brevith, they are not presented here, but are available upon request.

In our two placebo tests, we use the same matched samples as in Section 5.1. We use a generator of the pseudo-random numbers from the uniform (0,1) distribution. In the first placebo test, denoted as Placebo test #1, we randomly assigned the low and high difference in searching algorithms for each firm and month. The high and low probability was set to 30%, consistent with Section 5.1. This approach, however, does not allow for any momentum in the attention of investors, since the distribution is randomly generated for each month. Consequently, we employ a second placebo test, denoted as Placebo test #2, that assigns, with the same probability, high and low attention in two subsequent periods, rather than generating every month individually. For the evaluation of the placebo treatment, we use the same approach as in Section 5.1. We report our results in Table 5.

------------------------------ Insert Table 5 around here ------------------------------

Similar to Table 4, we use two different measures for effective spread. Column (1) and (3) use dollar-weighted effective spread scaled by midquote, while columns (2) and (4) use the share-weighted effective spread. We see for both Placebo tests #1 and #2, the treatment effect is not significant. This rejects the concern that the SVI difference is randomly generated. These findings support our identification strategy and verify that our results are driven by technological improvements, which implies the attention of unsophisticated or uninformed investors.

6. Summary

In this study, we use a natural experiment based on technological improvements to Google Trends data to provide a new measure for the attention of unsophisticated or uninformed investors. We find that unlike the attention of informed investors, which has a positive effect on liquidity, the attention of unsophisticated or uninformed investors has a negative effect on liquidity. This result explains the mixed evidence reported in the literature regarding the liquidity provision hypothesis. The negative effect on liquidity extends into stock performance, which in turn affects the firm's cost of capital. We find, however, that these effects are limited to smaller firms with lower levels of institutional equith investment. For larger firms, institutional investors offset any adverse effects from unsophisticated or uninformed investors traders. We find that their participation has a positive effect on liquidity and reduces the firm's cost of capital. Our methodology is validated using a set of placebo tests. We use the Average Treatment Effect on Treated (ATET) to confirm the causal relationship.

While our sample terminates in 2018, our results can help to explain more recent events, such as the GameStop shorts queeze and the Meme Stock mania. Given the surge in retail investor participation in the stock market, the question remains if these new investors are as I tgheir nformed as established investors. It is increasingly likely that these new investors who are attracted by gamified platforms and the fear of missing out, are less knowledgeable and are less informed. This study shows that such investors are potentially harmful to market liquidity and make it more difficult for smaller firms to raise capital.

These results are useful to a number of audiences and communities. By more correctly assessing the impact of retail investors on market behaviour, more effective regulatory policies by agencies such as the SEC or the exchanges themselves can be created. Executives and managers of smaller firms can better understand the forces affecting their cost of investment capital and suggest strategies for securing funds at lower rates. Finally, investors themselves can gain valuable insights into the effect that even uninformed trading can have on equity prices.

References

Amihud Y., 2002, Illiquidity and stock returns: cross-section and time-series effects, *Journal of Financial Markets* 5, 31-56.

Armstrong Ch. S., Core J. E., Taylor D. J., Verrecchia R. E., 2010, When Does Information Asymmetry Affect the Cost of Capital? *Journal of Accounting Research* 49(1), 1-40.

Barber B. M., Huang X., Jorion P., Odean T., Schwarz Ch., 2023, A (Sub)penny For Your Thoughts: Tracking Retail Investor Activity in TAQ, *Journal of Finance* forthcoming, Available at SSRN: https://ssrn.com/abstract=4202874 or http://dx.doi.org/10.2139/ssrn.4202874

Barber B. M., Odean T., Zhu N., 2009, Do Retail Trades Move Markets? *Review of Financial Studies* 22, 151-186.

Barrot J.-N., Kaniel R., Sraer D., 2016, Are retail traders compensated for providing liquidity? *Journal of Financial Economics* 120, 146-168.

Boehmer E., Jones Ch. M., Zhang X., Zhang X., 2021, Tracking Retail Investor Activity, *The Journal of Finance* 76(5), 2249-2305.

Black F., 1986, Noise, *The Journal of Finance* 41(3), 528-543.

Bradley D., Hanousek J., Jame R., Xiao Z., 2023, Place your bets? The market consequences of investment advice on Reddit's Wallstreetbets, *Review of Financial Studies* forthcoming, Available at SSRN: https://ssrn.com/abstract=3806065 or http://dx.doi.org/10.2139/ssrn.3806065.

Brennan M. J, Chordia T., Subrahmanyam A., Qing T., 2012, Sell-order liquidity and the cross-section of expected stock returns, *Journal of Financial Economics* 105(3), 523-541.

Chapkovski P., Khapko M., Zoican M., 2023, Trading gamification and investor behavior. Management Science (forthcoming), Available at SSRN: https://ssrn.com/abstract=3971868 or http://dx.doi.org/10.2139/ssrn.3971868

Chordia, T., Roll, R., Subrahmanyam, A., 2001. Market liquidity and trading activity. *The Journal of Finance* 56, 501–530.

Da, Z., Engelberg, J., Gao, P., 2011, In search of attention, *Journal of Finance* 66 (5), 1461–1499.

Diamond D. W., Verrecchia R. E., 1991, Disclosure, Liquidity, and the Cost of Capital, *The Journal of Finance* 46(4), 1325-1359.

Ding R., Hou W., 2015, Retail investor attention and stock liquidity, *Journal of International Financial Markets, Institutions and Money* 37, 12-26.

Eichenauer V. Z., Indergand R., Martínez I. Z., Sax Ch., 2021, Obtaining consistent time series from Google Trends, *Economic Inquiry* 60, 694-705.

Fama E. F., French K. R., 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116(1), 1-22.

Fang V. W., Noe T. H., Tice S., 2009, Stock market liquidity and firm value, *Journal of Financial Economics* 94, 150-169.

Farrell M., Green T. C., Jame R., Markov S., 2022, The democratization of investment research and the informativeness of retail investor trading, *Journal of Financial Economics* 145(2), 616-641.

Frazzini A., Lamont O. A., 2008, Dumb money: Mutual fund flows and the cross-section of stock returns, *Journal of Financial Economics* 88, 299-322.

Goldstein I., Ozdenoren E., Yuan K., 2013, Trading frenzies and their impact on real investment, *Journal of Financial Economics* 109, 566-582.

Green, J., Smart, S., 1999, Liquidity Provision and Noise Trading: Evidence from the "Investment Dartboard" Column, *The Journal of Finance* 54(5), 1885-1899.

Grossman S. J., Stiglitz J. E., 1980, On the Impossibility of Informationally Efficient Markets, *The American Economic Review* 70(3), 393-408.

Hirshleifer D., Sonya S. L., Teoh S. H., 2011, Limited Investor Attention and Stock Market Misreactions to Accounting Information, The Review of Asset Pricing Studies 1(1), 35-73.

Hirshleifer D., Teoh S. H., 2003, Limited attention, information disclosure, and financial reporting, *Journal of Accounting Economics* 36(1-3), 337-386.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945-960.

Hou, K., Peng, L., Xiong, W., 2009, A tale of two anomalies: The implications of investor attention for price and earnings momentum, Working paper, Ohio State University and Princeton University.

Hvidkjaer S., 2008, Small Trades and the Cross-Section of Stock Returns, *Review of Financial Studies* 21, 1123-1151.

Kaniel R., Saar G., Titman S., 2008. Individual Investor Trading and Stock Returns, *The Journal of Finance* 68(1), 273-310.

Kaniel R., Liu S., Saar G., Titman S., 2012, Individual Investor Trading and Return Patterns around Earnings Announcements, *The Journal of Finance* 67, 639-680.

Kelley E. K., Tetlock P. C., 2013, How Wise Are Crowds? Insights from Retail Orders and Stock Returns, *The Journal of Finance* 68(3), 1229-1265.

Kumar A., Lee CH. M.C., 2006. Retail Investor Sentiment and Return Comovements, *The Journal of Finance* 61(5), 2451-2486.

Lee C. M. C., 2001, Market efficiency and accounting research: A discussion of 'capital market research in accounting' by S.P. Kothari, *Journal of Accounting and Economics* 31(1-3), 233–253.

Lee C. M. C., Ready M. J., 1991, Inferring Trade Direction from Intraday Data, *The Journal of Finance* 46(2), 733-746.

McCabe C., 2021, Individual Investors Retreat from Markets after Show-Stopping Start to 2021, *Wall Street Journal* (April 4).

O'Brien, P. C., & Bhushan, R. (1990). Analyst following and institutional ownership. *Journal of Accounting Research*, 28, 55-76.

Ozik G., Sadka R., Shen S., 2021, Flattening the Illiquidity Curve: Retail Trading During the COVID-19 Lockdown, *Journal of Financial and Quantitative Analysis* 56(7), 2356-2388.

Rosenbaum, P. R., and D. B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 39: 33–38. https://doi.org/10.2307/2683903.

Rubin, D. B. 2008. For objective causal inference, design trumps analysis. *Annals of Applied Statistics* 2: 808–840. https://doi.org/10.1214/08-AOAS187.

Shiller, Robert J., 1984, Stock prices and social dynamics, *Brookings Papers on Economic Activity* 2, 457–510.

Shleifer A., 2000, Inefficient Markets – An Introduction to Behavioral Finance, *Oxford University Press*, Oxford, UK.

Shleifer A., Summers L. H., 1990, The Noise Trader Approach to Finance, *Journal of Economic Perspectives* 4(2), 19-33.

Shleifer A., Vishny R. W., 1997, The Limits of Arbitrage, *The Journal of Finance* 52(1), 35-55.

Figure 1 – Search Volume Index (SVI) of the ticker of Google (“GOOG”)

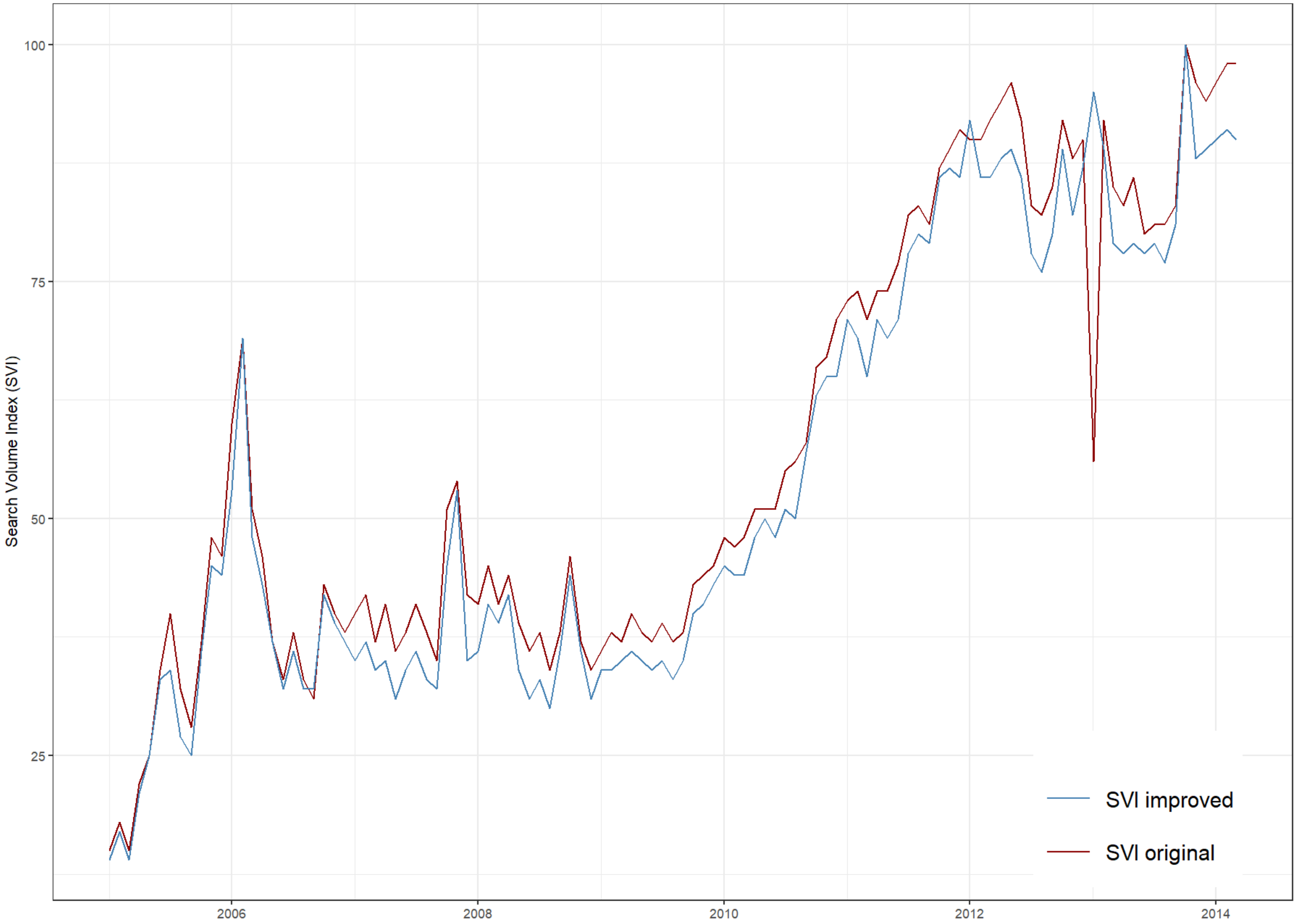

Table 1: Investor attention and retail volume

This table reports the effects of investor attention on retail volume. To capture retail volume, we use the Barber et al. (2023) improvement on the Boehmer et al. (2021) algorithm, which uses the Lee and Ready (1991) quote midpoint signing method. Columns 1 to 3 used as a dependent variable retail volume scaled by the total volume for the month. Columns 4 to 6 used as a dependent variable the abnormal retail volume, which is the percentage change of retail volume to the average retail volume for the stock. The average retail volume is calculated by taking the average retail volume for the stock over the past three months. Control variables for every regression include Size, Book to market, Past profitability, and Amihud's illiquidity, including the year and firm dummies. Standard errors are clustered at the firm level to control for unobserved time-invariant firm-level heterogeneity. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

	Retail volume scaled			**Abnormal retail volume**		
Variables	**Model (1)**	**Model (2)**	**Model (3)**	**Model (4)**	**Model (5)**	**Model (6)**
SVI original	0.0206***			0.0097***		
	(0.0015)			(0.0012)		
SVI improved		0.0222***			0.0102***	
		(0.0018)			(0.0016)	
SVI difference			0.0101***			0.0057***
			(0.0015)			(0.0006)
Size	-1.7980***	-1.7902***	-1.7476***	-0.0920***	-0.0911***	-0.0721***
	(0.0784)	(0.0791)	(0.0805)	(0.0248)	(0.0253)	(0.0246)
Book to market	0.5115***	0.5148***	0.5394***	-0.0703***	-0.0751***	-0.0641***
	(0.0670)	(0.0720)	(0.0740)	(0.0216)	(0.0227)	(0.0213)
R_{m-1}	-0.2331***	-0.2347***	-0.2149***	0.0257	0.0255	0.0345
	(0.0555)	(0.0556)	(0.0560)	(0.0320)	(0.0322)	(0.0320)
$R_{[m-3,m-2]}$	0.7884***	0.7923***	0.8155***	-0.0942***	-0.0935***	-0.0831***
	(0.0679)	(0.0680)	(0.0686)	(0.0319)	(0.0320)	(0.0312)
$R_{[m-6,m-4]}$	0.7268***	0.7240***	0.7317***	-0.0705***	-0.0717***	-0.0682***
	(0.0655)	(0.0655)	(0.0662)	(0.0158)	(0.0159)	(0.0154)
$R_{[m-12,m-7]}$	0.1692***	0.1692***	0.1742***	-0.0243***	-0.0251***	-0.0228**
	(0.0315)	(0.0315)	(0.0318)	(0.0090)	(0.0091)	(0.0089)
Illiquidity	-0.0004	-0.0004	-0.0003	-0.0011	-0.0011	-0.0011
	(0.0011)	(0.0012)	(0.0011)	(0.0007)	(0.0007)	(0.0007)
Constant	14.9924***	14.9145***	15.3370***	0.5589***	0.5733***	0.7669***
	(0.5499)	(0.5588)	(0.5687)	(0.1893)	(0.1916)	(0.1957)
Firm fixed effects	YES	YES	YES	YES	YES	YES
Year fixed effects	YES	YES	YES	YES	YES	YES
R^2	0.6972	0.6916	0.6895	0.0375	0.0368	0.0336
Number of observations	264,878	263,770	263770	264,840	263,732	263,732

Table 2: The Effect of Investor Attention on the Effective Spread

This table reports the effects of investor attention on the effective spread. The dependent variable is the natural logarithm of the dollar-weighted effective spread scaled by midquote. The dependent variable is negatively related to liquidity. Results are identical when using Share-weighted effective spread. Control variables for every regression include Size, Book to market, Past profitability, and Amihud's illiquidity, including the year and firm dummies. Standard errors are clustered at the firm level to control for unobserved time-invariant firm-level heterogeneity. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

Variables	**Model (1)**	**Model (2)**	**Model (3)**	**Model (4)**
SVI original	-0.0019***			
	(0.0003)			
SVI improved		-0.0019***		
		(0.0002)		
SVI difference			0.0012***	0.0079***
			(0.0004)	(0.0008)
SVI difference*Q2 size				-0.0091***
				(0.0010)
SVI difference*Q3 size				-0.0116***
				(0.0010)
SVI difference*Q4 size				-0.0062***
				(0.0013)
Size	-0.5407***	-0.5394***	-0.5454***	-0.5383***
	(0.0064)	(0.0065)	(0.0065)	(0.0065)
Book to market	-0.0063	-0.0029	-0.0066	-0.0095
	(0.0111)	(0.0117)	(0.0120)	(0.0117)
R_{m-1}	-0.1890***	-0.1889***	-0.1909***	-0.1895***
	(0.0068)	(0.0068)	(0.0069)	(0.0070)
$R_{[m-3,m-2]}$	-0.1902***	-0.1907***	-0.1910***	-0.1832***
	(0.0087)	(0.0088)	(0.0088)	(0.0087)
$R_{[m-6,m-4]}$	-0.0036	-0.0038	-0.0025	0.0015
	(0.0084)	(0.0084)	(0.0085)	(0.0083)
$R_{[m-12,m-7]}$	0.0116	0.0114	0.0117	0.0128*
	(0.0076)	(0.0075)	(0.0075)	(0.0078)
Illiquidity	0.0004**	0.0004**	0.0004**	0.0036**
	(0.0002)	(0.0002)	(0.0002)	(0.0015)
Constant	-1.9956***	-2.0069***	-2.0553***	-2.1215***
	(0.2099)	(0.2116)	(0.2050)	(0.1994)
Industry fixed effects	YES	YES	YES	YES
Year fixed effects	YES	YES	YES	YES
R^2	0.7706	0.7717	0.7698	0.7724
Number of observations	267,583	266,475	266,475	265,899

Table 3: Unsophisticated investor attention and cost of capital

This table reports the effect of unsophisticated retail investors' attention on the cost of capital. We form 25 equal-weighted portfolios for each month based on two-dimensional dependent sorts and compute one-month ahead buy-and-hold returns for each portfolio. Firms are first ranked into quintiles based on unsophisticated retail investor attention, which is the difference between the original SVI value and the improved SVI value. Then, within each number of unsophisticated retail investors' attention, they are further sorted into five portfolios based on size, defined as the natural logarithm of the market value of equity. We also include an arbitrage portfolio created by buying the largest firms and selling the smallest ones within the given pentile of unsophisticated investor attention. The portfolio is rebalanced every month from January 2010 until December 2018. We use the Fama-French 5-factor model:

$$R_m^P - R_m^F = \beta_0 + \beta_1(R_m^M - R_m^F) + \beta_2 SMB_m + \beta_3 HML_m + \beta_4 RMW_m + \beta_5 CMA_m + \epsilon_m$$

where R_m^P is the monthly return of a particular portfolio, R_m^F is the one-month Treasury bill rate, and R_m^M is the value-weighted market return. ***, **, and * denote statistical significance at 1%, 5%, and 10% levels, respectively.

Panel A: Smallest size pentile

Variables	Q1 (Smallest attention)	Q5 (Largest attention)	Arbitrage portfolio (smallest - largest)
$R_m^M - R_m^F$	0.9724***	0.9159***	0.0571
	(0.0628)	(0.0726)	(0.0664)
SMB	0.8855***	0.7474***	0.1397
	(0.0969)	(0.1120)	(0.1025)
HML	0.1785	0.0601	0.1208
	(0.1246)	(0.1440)	(0.1318)
RMW	-0.4213***	-0.4874***	0.0688
	(0.1522)	(0.1760)	(0.1610)
CMA	-0.0100	0.1738	-0.1855
	(0.1850)	(0.2139)	(0.1957)
Constant	-1.7173***	-2.3357***	0.5917**
	(0.2158)	(0.2495)	(0.2282)
R^2	0.8610	0.7970	0.0487
N (observations)	108	108	108

Panel B: Largest size pentile

Variables	**Q1 (Smallest attention)**	**Q5 (Largest attention)**	**Arbitrage portfolio (smallest - largest)**
$R_m^M - R_m^F$	1.0104***	0.9977***	0.0132
	(0.0229)	(0.0276)	(0.0304)
SMB	0.1129***	0.4487***	-0.3343***
	(0.0354)	(0.0426)	(0.0468)
HML	0.0102	-0.1099**	0.1226**
	(0.0455)	(0.0548)	(0.0602)
RMW	0.0218	-0.1627**	0.1872**
	(0.0556)	(0.0670)	(0.0736)
CMA	-0.1106	-0.0588	-0.0536
	(0.0676)	(0.0814)	(0.0895)
Constant	0.4000***	0.8110***	-0.4377***
	(0.0788)	(0.0949)	(0.1043)
R^2	0.9609	0.9552	0.4610
N (observations)	108	108	108

Table 4: Effect of unsophisticated investor's attention on effective spread. RCT approach.

This table reports the Average Treatment Effect on Treated (ATETT), measuring the impact of the increased attention of unsophisticated investors on the effective spread. We use the distribution of the SVI difference, where we assign monthly values above the 70th percentile as the treatment group (attention driven by the unsophisticated investors) and values below the 30th percentile serve a control group. We require exact matching on year, month, and Fama-French 48 industry classification, while the approximate coordinate for matching is the firm size. For Panel A, column (1) uses Dollar-weighted effective spread scaled by midquote, and column (2) uses Share-weighted effective spread scaled by midquote.

In each column, we report the ATET conducted as an effect of the high attention of the unsophisticated investors. The standard errors of the ATET (in parentheses) are computed with the robust option (at least two suitable matches for each treated). Below is the balance summary of the mean difference and variance ratio between the corresponding treated and control groups. Panel B shows ther density plot outlining matching quality. ***, **, and * denote statistical significance on 1, 5, and 10% significance levels.

Panel A – Results of ATET analysis

RCT output	Mean effective spread	
	(1)	(2)
High attention unsophisticated (ATET)	0.0833***	0.0834***
(std. error)	(0.005)	(0.005)
p-value	<0.001	<0.001
Number of treated	113,531	113,531
Number of observations	231,626	231,626
Balance summary		
mean difference (size)	0.001	0.001
variance ratio (size)	1.013	1.013

Panel B – Density plot of matching quality

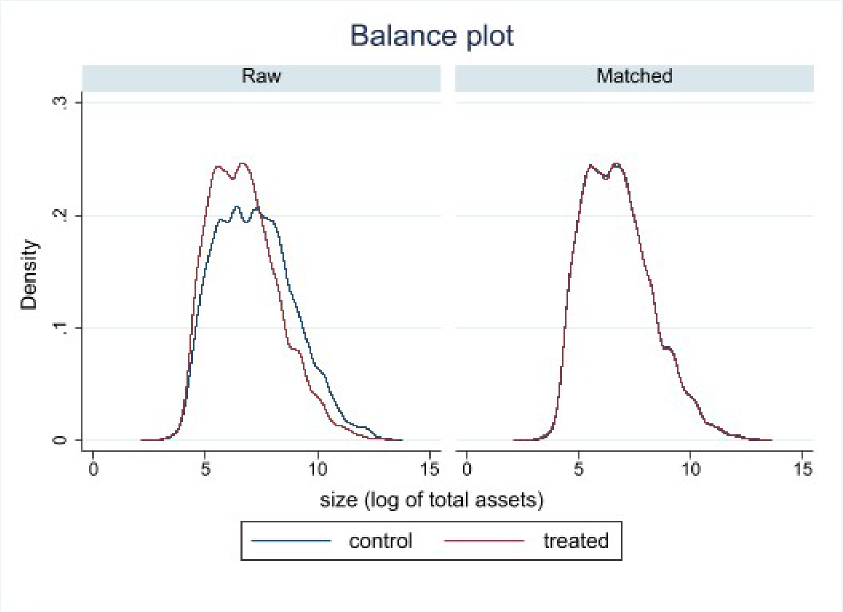

Table 5: Placebo test

This table reports the Average Treatment Effect on Treated (ATET), measuring the impact of the randomly assigned increased search by unsophisticated investors. We use a generator of the pseudo-random numbers from the uniform (0,1) distribution. For replication purposes, we present below the results with the random seed number equal to 12345. The Placebo test #1 randomly for each firm and month assigned the low and high difference in searching algorithms (linked with the unsophisticated investors). The probability of high and low was set to 30%. Before we require the same matching of the placebo assignments, we exclude about 40% of firms in the middle. Placebo test #2 assigns (with the same probability) the high and low in two subsequent periods. For the evaluation of the placebo treatment, we use the same set of covariates and require the same matching as for the real identification (year, month, and Fama-French 48 industry classification).

Columns (1) and (3)) uses Dollar-weighted effective spread scaled by midquote, and columns (2) and (4) correspond to the variable Share-weighted effective spread scaled by midquote

In each column, we report the placebo test conducted as the mean effect on treated (ATET), the difference from the similar firms in high and low (unsophisticated vs. the rest). The standard errors of the ATET (in parentheses) are computed with the robust option (at least two suitable matches for each treated).

Below is the balance summary of the mean difference and variance ratio between the corresponding treated and control groups.

RCT output	Placebo test #1		Placebo test #2	
	(1)	(2)	(3)	(4)
Test value (ATET)	-0.0005	-0.0004	-0.0045	-0.0048
(std. error)	(0.006)	(0.006)	(0.005)	(0.006)
p-value	0.94	0.95	0.40	0.38
Number of treated	75,954	75,954	89,153	89,153
Number of observations	151,856	151,856	177,942	177,942
Balance summary				
mean difference (size)	0.013	0.013	0.013	0.013
variance ratio (size)	1.096	1.096	1.096	1.096

Appendix A – Variable descriptions

Variable	*Description*
Dependent variables	
Retail volume scaled	To capture retail volume, we use the Barber et al. (2023) improvement on the Boehmer et al. (2021) algorithm, which uses the Lee and Ready (1991) quote midpoint signing method. We then define Retail volume scaled as the retail volume for the month scaled by the total volume for the month. Data source: TAQ
Abnormal retail volume	To capture retail volume, we use the Barber et al. (2023) improvement on the Boehmer et al. (2021) algorithm, which uses the Lee and Ready (1991) quote midpoint signing method. Abnormal retail volume is the percentage change of retail volume to the average retail volume for the stock. The average retail volume is calculated by taking the average retail volume for the stock over the past three months. Data source: TAQ
Dollar-weighted effective spread scaled by midquote	Natural logarithm of dollar-weighted effective spread scaled by the midquote. The effective spread is calculated at daily frequency and then we take the average during the given calendar month. Source: TAQ
Share-weighted effective spread scaled by midquote	Natural logarithm of share-weighted effective spread scaled by the midquote. The effective spread is calculated at daily frequency and then we take the average during the given calendar month. Source: TAQ
Measures of investor attention	
SVI original	Search Value Index available through Google Trends. Index has a range from 0 to 100, and is scaled by the maximum value in the series. SVI original was collected in December 2019.
SVI improved	Search Value Index available through Google Trends. Index has a range from 0 to 100, and is scaled by the maximum value in the series. SVI improved was collected in December 2022.
SVI difference	Difference between SVI original and SVI improved
Firm control variables	
Firm size	Firm size is the natural logarithm of the market value of equity. Data sources: CRSP and Compustat.
Book-to-market ratio	The book-to-market ratio is defined as book equity divided by market equity. Data sources: CRSP and Compustat.
Past profitability	The group of variables R_{m-1}, $R_{[m-3,m-2]}$, $R_{[m-6,m-4]}$, and $R_{[m-12,m-6]}$, which stand for returns over the last month, months 3 to 2, 6 to 4, and 12 to 6, respectively. Defined by Brennan et al. (2012). Data sources: CRSP and Compustat.
Illiquidity	Illiquidity is the sum of the absolute values of daily returns divided by the daily volume for the year, multiplied by 10^6. Defined by Amihud (2002). Data sources: CRSP and Compustat.

Appendix B – Summary statistics

	N	Mean	SD	P25	Median	P75
Measures of investor attention						
SVI original	288,793	41.261	25.124	19.000	39.000	61.000
SVI improved	287,636	37.818	26.474	14.000	35.000	59.000
SVI difference	287,636	3.488	9.789	-1.000	1.000	6.000
Retail volume						
Retail volume scaled	285634	5.994	4.983	2.884	4.263	7.171
Abnormal retail volume	285456	0.133	1.807	-0.294	-0.064	0.260
Effective spread						
Dollar-weighted scaled by midquote	288,725	-3.341	0.980	-4.120	-3.474	-2.754
Share-weighted scaled by midquote	288,725	-3.341	0.979	-4.120	-3.474	-2.753
Firm Characteristics						
Size	284,308	7.035	1.692	5.717	6.876	8.129
Book to market	274,091	0.670	19.765	0.278	0.511	0.817
Illiquidity	288,723	0.396	40.756	0.000	0.002	0.015
R_{m-1}	288,709	1.012	0.127	0.952	1.009	1.066
$R_{[m-3,m-2]}$	288,536	1.025	0.179	0.937	1.019	1.101
$R_{[m-6,m-4]}$	287,373	1.041	0.223	0.930	1.030	1.135
$R_{[m-12,m-6]}$	282,125	1.111	0.459	0.922	1.068	1.228